MatchSense: Multimodal Reasoning for Key Moment Localization in Long-Form Sports Broadcasts

Anonymous ACL submission

Abstract

Long-form sports broadcasts contain sparse but decisive events within extended background play, making key moment identification difficult for current video–language models. We introduce **MatchSense**, an audio-guided language reasoning framework for identifying key moments in soccer broadcasts. MatchSense first uses prosodic saliency to extract Salient Object Intervals (SOIs), reducing the temporal search space before applying structured video–language reasoning with <think> and <answer> outputs. To handle variable event counts and heterogeneous reward signals, we adapt Group reward-Decoupled Normalization Policy Optimization (GDPO) to assignment-based key moment prediction. Predicted and ground-truth events are first matched through Hungarian assignment, after which temporal overlap, semantic alignment, output-format validity, and temporally aligned acoustic saliency are normalized as separate reward components before policy optimization. Under our video–language key moment identification setting, MatchSense improves mAP to 9.42%, reduces mean boundary error to 1.54s, and improves Precision@0.5 from 6.80% to 11.50% over SFT-only. Audio pruning retains 35.3% of the timeline while preserving 83% of salient events, yielding a 64.7% FLOP reduction and 2.8× speedup. All code and implementations are publicly available via an anonymous repository at https://anonymous.4open.science/r/MatchSense-6FE4/README.md.

1 Introduction

Long-form sports broadcasts contain sparse but decisive moments within extended background play. In soccer, goals, penalties, fouls, shots, cards, substitutions, and set pieces shape the match narrative, yet occupy only a small fraction of a 90-minute video. Identifying these moments supports highlight generation, retrieval, summarization, commentary analysis, and decision support, but remains challenging due to long context, sparse annotations, ambiguous boundaries, and visually repetitive play.

We study *Key Moment Identification in Long-Form Sports Broadcasts*: localizing and labeling semantically consequential soccer events from untrimmed matches. Unlike generic temporal grounding, the target moments are broadcast-relevant events requiring multimodal language reasoning over visual action, commentator speech, crowd reaction, and acoustic saliency.

We propose **MatchSense**, an audio-guided language reasoning framework. MatchSense extracts *Salient Object Intervals* (SOIs) using prosodic cues such as energy, onset activity, and speech-related variation. This retains 35.3% of the timeline, averaging 31.8 minutes per 90-minute match, while preserving 83% of salient events, reducing FLOPs by 64.7% and yielding a 2.8× speedup over dense processing.

Within SOIs, MatchSense aligns SoccerNet-v2 annotations with fixed-length 30-second clips and trains a video–language model to generate <think> descriptions followed by <answer> event labels and intervals. SFT improves mAP from 1.25% to 5.85%, but still yields 2.45s mean boundary error and only 2.06% F1@0.7. We therefore adapt Group reward-Decoupled Normalization Policy Optimization (GDPO) (Liu et al., 2026) to assignment-based key moment prediction. Hungarian matching compares unordered prediction sets with ground-truth events, while GDPO separately normalizes temporal, semantic, acoustic, format, and penalty rewards before aggregation. The full model reaches 9.42% mAP, reduces mean boundary error to 1.54s, improves Precision@0.5 from 6.80% to 11.50%, and raises information density from 7.1% to 16.7%.

Contributions. We make four primary contributions: **(1) Video Filtering:** We propose **MatchSense**, which leverages prosodic saliency to condense long-form broadcasts, retaining 35.3% of the timeline while preserving 83% of salient events. **(2)**

Structured Reasoning: We induce interpretable reasoning via <think> and <answer> outputs, where the <think> tag grounds predictions by generating timestamp-wise natural language descriptions of game play. **(3) Hungarian-Matched GDPO:** We adapt GDPO for variable-cardinality prediction using Hungarian matching, optimized via a decoupled reward function encompassing format compliance, semantic alignment, acoustic prosody, temporal accuracy, and strict penalties. **(4) Empirical Validation:** We demonstrate robust improvements across mAP, boundary precision, Precision@0.5, information density, and inference efficiency.

2 Related Work

Language-grounded video moment retrieval. Temporal sentence grounding, or video moment retrieval, localizes video segments corresponding to natural-language queries. Early methods such as MCN and TALL aligned visual–textual representations or regressed temporal boundaries for language queries (Anne Hendricks et al., 2017; Gao et al., 2017). Later benchmarks such as ActivityNet Captions, TVR, and MAD introduced denser and longer multimodal grounding settings (Krishna et al., 2017; Lei et al., 2020; Soldan et al., 2022). However, most VMR tasks remain query-conditioned and clip-centric, whereas our setting requires identifying multiple broadcast-relevant key moments from full-length sports videos without an explicit user query.

Sports action spotting and long-form grounding. Sports video understanding has been studied through action spotting, temporal localization, and highlight detection. SoccerNet established full-match soccer analysis as sparse event spotting over long broadcasts (Giancola et al., 2018; Deliege et al., 2021), while BMN and ActionFormer model temporal boundaries in untrimmed videos (Lin et al., 2019; Zhang et al., 2022). Recent long-video models such as ReVisionLLM and Temporal Chain of Thought extend grounding through sparse-to-dense search or compact motion representations (Hannan et al., 2025; Arnab et al., 2026). In contrast, MatchSense reduces irrelevant background processing by pruning the search space before video–language reasoning.

Audio, prosody, and language reasoning. Audio cues such as commentator excitement, crowd noise, acoustic onsets, pitch, and energy often correlate with salient sports events (Xiong et al., 2003). Prior multimedia work uses audio and prosody for highlight detection and affective video understanding (Merler et al., 2019), consistent with computational paralinguistics, where F0, energy, and intensity indicate arousal (Schuller et al., 2013; El Ayadi et al., 2011). MatchSense uses these cues to construct Salient Object Intervals and then applies structured <think>–<answer> reasoning to connect visual dynamics, commentary-derived semantics, and event localization.

Preference optimization and assignment-based localization. Supervised grounding objectives often do not directly optimize temporal overlap. Time-R1 addresses this using reinforcement learning with verifiable overlap-aware rewards (Wang et al., 2026; Shao et al., 2024). However, sports broadcasts contain variable event counts, nearby events, and duplicate predictions from dense windows. We therefore use Hungarian assignment to match predicted and ground-truth key moments before computing temporal, semantic, and acoustic rewards (Kuhn, 1955; Carion et al., 2020).

Research Gap. Prior work separately studies language-grounded retrieval (Gao et al., 2017; Anne Hendricks et al., 2017; Lei et al., 2020; Soldan et al., 2022), sports action spotting (Giancola et al., 2018; Deliege et al., 2021; Zhang et al., 2022), long-form grounding (Hannan et al., 2025), and RL-based temporal localization (Wang et al., 2026). Yet full-length sports broadcasts require a unified treatment of sparse event discovery, multimodal saliency, structured language reasoning, and variable-cardinality localization. Existing retrieval methods are largely query-conditioned, action-spotting systems focus on coordinate-level detection, long-form grounding models often lack broadcast-specific saliency priors, and RL-based localization does not directly handle unordered sports event sets. MatchSense fills this gap through prosodic pruning, <think>–<answer> reasoning, and Hungarian-matched multimodal reward optimization.

3 Methodology

We formulate *Key Moment Identification in Long-Form Sports Broadcasts* as structured multimodal set prediction. Given a full-length soccer broadcast, the goal is to localize and label semantically important events under extreme temporal sparsity.

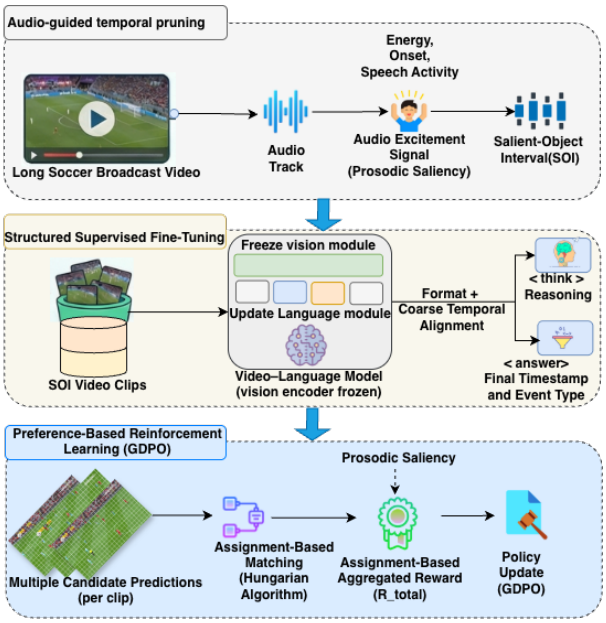

Figure 1: Overview of **MatchSense**. A full-length soccer broadcast is pruned into Salient Object Intervals (SOIs) using audio prosody. SOI clips are then used for structured supervised fine-tuning with <think> and <answer> outputs, followed by assignment-based Group reward-Decoupled Normalization Policy Optimization (GDPO).

MatchSense addresses this through three stages: audio-driven candidate generation, structured supervised fine-tuning, and assignment-based GDPO.

3.1 Task Formulation

Let V be a broadcast of duration T with timestamped annotations $\mathcal{E} = \{(y_i, \tau_i)\}_{i=1}^{N}$, where $y_i \in \mathcal{Y}$ is a soccer event label and $\tau_i \in [0, T]$ is its timestamp. Since SoccerNet-style annotations are point-level, each timestamp is expanded into an interval: $\tau_i \mapsto [\tau_i - \delta, \tau_i + \delta]$, yielding ground-truth key moments $\mathcal{G} = \{(y_i, [s_i, e_i])\}_{i=1}^{N}$. The model learns a function f_θ that predicts an unordered set: $f_\theta(V) = \{(\hat{y}_j, [\hat{s}_j, \hat{e}_j])\}_{j=1}^{M}$, where M is unknown, $\hat{y}_j \in \mathcal{Y}$, and $0 \le \hat{s}_j \le \hat{e}_j \le T$. Unlike single-query moment retrieval, the model must identify multiple events, handle variable cardinality, and suppress false positives over long background regions where $N \ll T$.

3.2 Audio-Driven Candidate Generation

To avoid exhaustive reasoning over the full timeline, MatchSense extracts *Salient Object Intervals* (SOIs) $\mathcal{S} = \{[a_k, b_k]\}_{k=1}^{K}$ from broadcast audio. The audio is converted to mono at 16 kHz, and short-time RMS energy, onset density, and an MFCC-derived speech-activity proxy are computed and normalized to $[0, 1]$. We construct an excitement signal: $E(t) = \alpha_1 E_{\text{energy}}(t) + \alpha_2 E_{\text{onset}}(t) + \alpha_3 E_{\text{speech}}(t), \quad \alpha_1 + \alpha_2 + \alpha_3 = 1$. After baseline subtraction and smoothing, salient peaks are selected using percentile, prominence, and separation constraints. Each peak t_p is expanded into $[a_k, b_k] = [t_p - \Delta_{\text{pre}}, t_p + \Delta_{\text{post}}]$, and overlapping intervals are merged. Events overlapping an SOI are converted into clip-relative coordinates by subtracting the SOI start time. Hyperparameters for SOI extraction are provided in Appendix A.8.

3.3 Structured SFT with Language Reasoning

For supervised fine-tuning, we construct 30-second clips bucketed by event count to cover zero-event, single-event, and multi-event cases. The model is trained to generate structured outputs: <think> ⋯ </think> <answer> ⋯ </answer>. The <think> block describes local play dynamics, while the <answer> block contains event labels and intervals. Frames are uniformly sampled, the vision encoder is frozen, and trainable adaptation components learn to map video evidence to structured language outputs. The SFT objective is $\max_\theta \sum_{(V_k, y_k) \in \mathcal{D}_{\text{SFT}}} \log p_\theta(y_k \mid V_k)$.

3.4 Assignment-Based Multi-Reward Scoring

SFT aligns the model with the output format but does not directly optimize temporal overlap or boundary precision. For each SOI clip, let $\mathcal{P}_k = \{p_j\}_{j=1}^{M_k}$ and $\mathcal{G}_k = \{g_i\}_{i=1}^{|\mathcal{G}_k|}$ denote predicted and ground-truth key moment sets. Since these sets are unordered and may have different cardinalities, we define a pairwise matching score:

$$m(p_j, g_i) = w_t r_{\text{tIoU}}(p_j, g_i) + w_s r_{\text{semantic}}(p_j, g_i) + w_a \big(r_{\text{tIoU}}(p_j, g_i) \cdot r_{\text{prosody}}(g_i)\big),$$

with $(w_t, w_s, w_a) = (0.6, 0.2, 0.2)$. The semantic reward is computed using BERTScore. For a matched pair (p_j, g_i), let z_j be the generated <think> description concatenated with the predicted label, and let c_i be the reference text formed from the ground-truth label and temporally aligned ASR commentary. We define

$$r_{\text{semantic}}(p_j, g_i) = \text{BERTScore}_{F_1}(z_j, c_i),$$

where token-level contextual embeddings are matched by maximum cosine similarity and aggregated using the standard BERTScore F_1 score (Zhang et al., 2019). The prosody term is scaled by

tIoU so that acoustic saliency contributes reward only when the prediction is temporally aligned with the salient event.

We treat $m(p_j, g_i)$ as a positive matching score. Since the Hungarian implementation is formulated as cost minimization, we pass the negated cost matrix $C_{j,i} = -m(p_j, g_i)$ to the solver and obtain the optimal matching π^*. From this matching, we compute reward components:

$$
\begin{aligned}
R_{\text{tIoU}} &= \sum_{(j,i)\in\pi^*} r_{\text{tIoU}}(p_j, g_i), \\
R_{\text{sem}} &= \sum_{(j,i)\in\pi^*} r_{\text{semantic}}(p_j, g_i), \\
R_{\text{aud}} &= \sum_{(j,i)\in\pi^*} r_{\text{tIoU}}(p_j, g_i) \cdot r_{\text{prosody}}(g_i), \\
R_{\text{fmt}} &= r_{\text{format}}(y), \\
R_{\text{pen}} &= -0.5N_{\text{unmatched}} - 1.0N_{\text{missed}}.
\end{aligned}
$$

Detailed reward grouping, weights, assignment visualization, and design rationale are provided in Appendix B.6 and Figure 9.

3.5 Group Reward-Decoupled Normalization Policy Optimization

Given an input x and a group of sampled outputs $\{y_j\}_{j=1}^{G}$, each output receives a reward vector

$$\mathbf{R}^{(j)} = [R_{\text{fmt}}^{(j)}, R_{\text{tIoU}}^{(j)}, R_{\text{sem}}^{(j)}, R_{\text{aud}}^{(j)}, R_{\text{pen}}^{(j)}].$$

Standard GRPO first aggregates these heterogeneous rewards and then normalizes the summed score. In contrast, GDPO normalizes each reward component separately within the rollout group:

$$A_q^{(j)} = \frac{R_q^{(j)} - \mu_q}{\sigma_q + \epsilon},$$

where μ_q and σ_q are computed over the group for reward component q. The normalized components are then weighted and aggregated into the final policy advantage. This prevents temporal, semantic, acoustic, format, and penalty signals from collapsing into a single normalized scalar. The full GDPO objective and comparison with GRPO are provided in Appendix B.5.

3.6 Inference and Timeline Reconstruction

At inference time, MatchSense extracts SOIs from the full broadcast and processes each SOI independently. Predicted clip-relative intervals are parsed from `<answer>` and mapped back to global time by adding the SOI start time. Duplicate predictions from overlapping SOIs are filtered by temporal overlap, yielding the final set of globally localized key moments.

4 Experimental Setup

We evaluate **MatchSense** for key moment identification in full-length SoccerNet-v2 soccer broadcasts.

Dataset. Experiments use full-length SoccerNet-v2 broadcasts. Each match contains two ∼45-minute halves with timestamped annotations from a soccer event ontology including *Goal*, *Corner*, *Foul*, and *Substitution*. Point annotations are expanded into short temporal windows to capture annotation uncertainty, event buildup, and immediate aftermath.

We extract Salient Object Intervals (SOIs) using the audio-driven candidate generation module in Section 3.2. SOIs provide variable-length event-centric clips for policy optimization and guide fixed-length 30-second clip construction for SFT. SFT clips are bucketed by overlapping event count to balance zero-event, single-event, and multi-event cases. For structured SFT, we construct a 200-clip training set from SoccerNet-v2 clips sampled across event-density buckets. This set is used to teach the model the required `<think>`–`<answer>` format and soccer event ontology. To verify format learning, we additionally evaluate a held-out 30-sample diagnostic set: before SFT, the base model produces valid outputs for 23/30 samples (76.67%), whereas after SFT it produces valid outputs for all 30 samples (100%). This diagnostic evaluates output-format adherence rather than final event localization accuracy. Dataset justification is provided in Appendix A.13.

Training. Our model was trained on an NVIDIA H100 80GB GPU for 224 hours over 2700 samples. We use two-stage adaptation: structured supervised fine-tuning followed by assignment-based multi-reward policy optimization.

Supervised fine-tuning. The base video–language model is fine-tuned on clips paired with structured outputs containing `<think>` reasoning and `<answer>` key moment predictions. Training uses causal language modeling to align the model with the soccer ontology, output format, and coarse temporal reasoning.

Assignment-based GDPO. After SFT, candidate outputs are parsed into key moment sets and matched to ground truth using Hungarian assignment. The matching induces multiple reward components, including output-format validity, temporal overlap, semantic alignment, temporally aligned acoustic saliency, and penalties for unmatched or missed events, as defined in Section 3.4. Our main model uses **GDPO**, i.e., Group reward-Decoupled Normalization Policy Optimization, which normalizes each reward component separately within a rollout group before aggregating the normalized advantages. We also report **GRPO** as a comparison variant, where reward components are first aggregated and then group-normalized. Further details are provided in Appendices B.4 and B.5.

Metrics. We report precision, recall, and F1 at tIoU thresholds 0.3, 0.5, and 0.7, together with mean Average Precision (mAP) and Mean Boundary Error (MBE, seconds). All metrics use optimal assignment between predicted and ground-truth events to handle variable prediction cardinality.

Baselines and implementation. We compare against zero-shot base models, SFT-only, Dense Sliding Window, GRPO/GDPO variants, reward ablations, no-`<think>` variants, and long-form grounding baselines including TGB and Time-R1-style variants. Where applicable, baselines use the same underlying video–language architecture, frame budget, and assignment-based evaluation protocol. Additional implementation and baseline adaptation details are provided in Appendix A.10.

5 Results

We evaluate **MatchSense** on key moment identification in full-length soccer broadcasts. We report event-level precision, recall, and F1 at temporal Intersection-over-Union (tIoU) thresholds $\{0.3, 0.5, 0.7\}$, together with mean Average Precision (mAP) and Mean Boundary Error (MBE, seconds). MBE is reported at tIoU 0.5 and measures boundary accuracy for successfully matched events. Unless otherwise stated, all predictions are matched to ground truth using one-to-one Hungarian assignment, ensuring consistent evaluation under variable prediction cardinality.

Scope of comparison. We do not claim state-of-the-art performance on SoccerNet action spotting. Specialized temporal spotters such as CALF, NetVLAD++, and ASTRA are optimized for coordinate-level action spotting and report substantially higher Average-mAP under SoccerNet protocols, e.g., NetVLAD++ reports 53.4% on SoccerNet-v2 and ASTRA reports 66.82% on the test set and 70.21% on the SoccerNet 2023 challenge set (Giancola and Ghanem, 2021). MatchSense targets a complementary setting: audio-guided pruning and structured video–language reasoning with interpretable `<think>`–`<answer>` outputs, trading raw timestamp-spotting precision for semantic reasoning and extensibility.

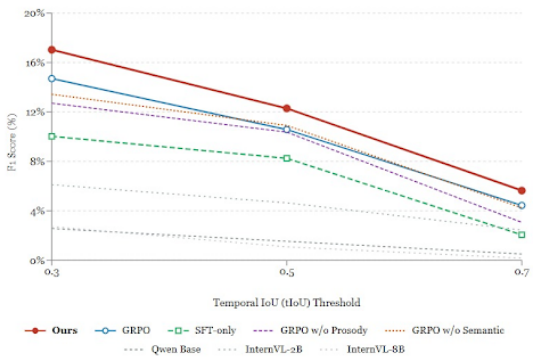

Figure 2: F1 score across tIoU thresholds $\{0.3, 0.5, 0.7\}$ on full-length SoccerNet-v2 broadcasts. MatchSense (GDPO) consistently achieves the highest F1 across thresholds, with the largest relative gains at stricter tIoU values.

5.1 Main Results

Table 1 compares zero-shot models, supervised fine-tuning, policy-optimized variants, and reward ablations. Three trends emerge. First, zero-shot video–language models struggle on sparse long-form sports events. Second, structured SFT improves event recognition and output formatting. Third, assignment-based multi-reward policy optimization improves boundary precision and event-level localization. Our main MatchSense model uses GDPO, i.e., Group reward-Decoupled Normalization Policy Optimization, which obtains the best overall performance by separately normalizing heterogeneous reward components before aggregation.

Figure 2 summarizes the threshold sensitivity of all methods. MatchSense maintains the strongest F1 at tIoU 0.3, 0.5, and 0.7, indicating that the improvement is not restricted to lenient matching. The relative gain is especially meaningful at tIoU 0.7, where stricter boundary agreement is required, supporting the claim that assignment-based GDPO improves temporal localization rather than only increasing event recall.

Table 1: Key moment identification performance on full-length soccer broadcasts. We report precision (P), recall (R), and F1 at temporal IoU thresholds $\{0.3, 0.5, 0.7\}$ together with mAP and Mean Boundary Error (MBE, seconds). All methods are evaluated using assignment-based matching. where InVL1: InternVL-2B; InVL2: InternVL-8B; QB: Qwen2.5-VL-7B Base; QSFT: Qwen2.5-VL-7B SFT-only; QGRPO: Qwen2.5-VL-7B GRPO; Ours: MatchSense (GDPO); G-NPR: GRPO w/o Prosody Reward; G-NSR: GRPO w/o Semantic Reward.

Method	**tIoU 0.3**			**tIoU 0.5**			**tIoU 0.7**			**mAP**	**MBE**
	P	R	F1	P	R	F1	P	R	F1		
InVL1	0.0374	0.1678	0.0612	0.0284	0.1275	0.0465	0.0150	0.0671	0.0245	0.0269	2.0432
InVL2	0.0158	0.1007	0.0273	0.0063	0.0403	0.0109	0.0011	0.0067	0.0018	0.0077	3.1733
QB	0.0208	0.0336	0.0257	0.0125	0.0201	0.0154	0.0042	0.0067	0.0051	0.0125	3.0370
QSFT	0.0825	0.1280	0.1003	0.0680	0.1050	0.0825	0.0180	0.0240	0.0206	0.0585	2.4500
QGRPO	0.1345	0.1620	0.1470	0.0980	0.1150	0.1058	0.0410	0.0485	0.0444	0.0815	1.6820
Ours	**0.1582**	**0.1845**	**0.1703**	**0.1150**	**0.1320**	**0.1229**	**0.0525**	**0.0610**	**0.0564**	**0.0942**	**1.5430**
G-NPR	0.1150	0.1420	0.1271	0.0965	0.1120	0.1037	0.0280	0.0340	0.0307	0.0712	2.1800
G-NSR	0.1210	0.1510	0.1343	0.1015	0.1180	0.1091	0.0395	0.0460	0.0425	0.0765	1.8450

Zero-shot models are unreliable for sparse broadcast events: InternVL-2B is the strongest zero-shot baseline, with 0.0269 mAP and 2.04s MBE, but its Precision@0.5 remains only 0.0284, indicating frequent hallucination in background play. InternVL-8B performs worse despite larger scale, with 0.0077 mAP and 3.17s MBE. Qwen2.5-VL-7B Base obtains 0.0125 mAP and 3.04s MBE, confirming that off-the-shelf video–language models are insufficient for precise key moment identification in full broadcasts.

SFT aligns format and ontology but not precise boundaries. Structured SFT improves Qwen2.5-VL-7B Base from 0.0125 to 0.0585 mAP and raises F1@0.5 from 0.0154 to 0.0825. This shows that SFT aligns the model with the soccer event ontology and `<think>`–`<answer>` output format. However, SFT-only still has 2.45s MBE and F1@0.7 of 0.0206, showing that next-token supervision does not sufficiently optimize temporal boundaries.

GDPO gives the strongest multi-reward optimization. Both GRPO and GDPO improve over SFT-only across all thresholds. GRPO reaches 0.0815 mAP and 1.68s MBE, while the main MatchSense model with GDPO performs best with 0.0942 mAP and 1.54s MBE. Compared with SFT-only, MatchSense improves Precision@0.5 from 0.0680 to 0.1150 and F1@0.5 from 0.0825 to 0.1229. This indicates that decoupled normalization of temporal, semantic, acoustic, format, and penalty rewards yields better localization than optimizing a single aggregated reward signal. Appendix B.7 further analyzes GDPO and GRPO training reward dynamics, showing that reward-decoupled normalization prevents early signal collapse under heterogeneous temporal, semantic, acoustic, format, and penalty rewards.

Table 2: Comparison with Dense Sliding Window inference. DSW evaluates the full broadcast uniformly, whereas MatchSense performs video–language reasoning only within audio-pruned intervals.

Method	**mAP@0.5**	**MBE(s)**
Qwen2.5-VL-7B Base	0.0125	3.037
Dense Sliding Window	0.0245	3.150
SFT-only + Audio Pruning	0.0585	2.450
MatchSense (GDPO)	**0.0942**	**1.543**

5.2 Effect of Audio-Guided Pruning

We compare MatchSense with Dense Sliding Window (DSW) inference to isolate the effect of audio-guided candidate generation. DSW partitions the full 90-minute broadcast into 30-second windows with 15-second stride, producing approximately 360 windows per match. The SFT model processes each window independently, and duplicate predictions are resolved using NMS with tIoU threshold 0.4.

DSW achieves high Recall@0.5 of 0.1480 because it evaluates the full timeline, but its Precision@0.5 is only 0.0210. Processing long stretches of uninformative play increases hallucinated detections, such as ordinary midfield passes being misclassified as fouls, clearances, or shots. Audio-pruned SFT improves Precision@0.5 to 0.0680, while MatchSense further improves it to 0.1150. DSW also has the worst boundary error, with 3.15s MBE. The same event may appear at different positions across overlapping win-

Table 3: Precision, recall, and F1 comparison between dense search and audio-pruned key moment identification. Where DSW is "Dense Sliding Window", SFT-AP is "SFT-only + Audio Pruning", and OUR's is "MatchSense (GDPO)".

Method	Metric	tIoU (0.3	0.5	0.7)
DSW	Precision	0.0310	0.0210	0.0050
	Recall	**0.1950**	**0.1480**	0.0320
	F1	0.0535	0.0367	0.0086
SFT-AP	Precision	0.0825	0.0680	0.0180
	Recall	0.1280	0.1050	0.0240
	F1	0.1003	0.0825	0.0206
OUR's	Precision	**0.1582**	**0.1150**	**0.0525**
	Recall	0.1845	0.1320	**0.0610**
	F1	**0.1703**	**0.1229**	**0.0564**

Table 4: Effect of structured language reasoning. Removing the <think> block degrades both supervised fine-tuning and policy optimization.

Method	mAP@0.5	MBE(s)
SFT w/o CoT	0.0450	2.850
SFT w/ CoT	0.0585	2.450
GRPO w/o CoT	0.0685	2.055
MatchSense (GDPO) w/ CoT	**0.0942**	**1.543**

dows, producing inconsistent local coordinates; NMS removes duplicates but can smear boundaries. MatchSense instead anchors candidate clips around acoustic peaks and optimizes boundaries through assignment-based multi-reward GDPO, reducing MBE to 1.54s.

5.3 Role of Structured Language Reasoning

We evaluate whether intermediate language reasoning helps key moment identification by comparing full models using <think> traces with variants trained without this reasoning block.

Removing the reasoning trace degrades performance at both stages. SFT without CoT obtains 0.0450 mAP, compared with 0.0585 for SFT with CoT, suggesting that directly mapping video evidence to event intervals is difficult. The <think> trace provides an intermediate linguistic representation of local play dynamics before boundary prediction. After policy optimization, GRPO without CoT improves to 0.0685 mAP but remains below MatchSense at 0.0942. Without generated reasoning text, the model cannot fully exploit the semantic reward component derived from commentary/event descriptions. The gap is largest at strict thresholds: F1@0.7 is 0.0228 without CoT versus 0.0564 for MatchSense, indicating improved fine-grained boundary awareness. A qualitative analysis of representative <think> traces, including correct causal localization, modality-induced hallucination, and missed-event cases, is provided in Appendix A.19.

Table 5: Detailed no-CoT comparison across temporal IoU thresholds. Our CoT model corresponds to **MatchSense (GDPO) w/ CoT**.

Method	Metric	tIoU (0.3	0.5	0.7)
SFT w/o CoT	Precision	0.0680	0.0580	0.0120
	Recall	0.1050	0.0655	0.0150
	F1	0.0825	0.0615	0.0133
GRPO w/o CoT	Precision	0.1040	0.0860	0.0210
	Recall	0.1380	0.0955	0.0250
	F1	0.1186	0.0905	0.0228
Ours w/ CoT	Precision	**0.1582**	**0.1150**	**0.0525**
	Recall	**0.1845**	**0.1320**	**0.0610**
	F1	**0.1703**	**0.1229**	**0.0564**

Table 6: Comparison with long-form grounding baselines. MatchSense improves both mAP and boundary precision by combining audio-guided pruning with assignment-based multi-reward GDPO.

Method	mAP@0.5	MBE(s)
Qwen2.5-VL-7B Base	0.0125	3.037
TGB	0.0415	2.650
Time-R1	0.0725	1.950
MatchSense (GDPO)	**0.0942**	**1.543**

Table 7: Detailed comparison with TGB and Time-R1 across tIoU thresholds. Where OUR's is "MatchSense (GDPO)"

Method	Metric	tIoU 0.3	tIoU 0.5	tIoU 0.7
TGB	Precision	0.0580	0.0450	0.0090
	Recall	**0.1850**	**0.1450**	0.0280
	F1	0.0883	0.0687	0.0136
Time-R1	Precision	0.1080	0.0860	0.0350
	Recall	0.1320	0.1040	0.0420
	F1	0.1189	0.0941	0.0382
OUR's	Precision	**0.1582**	**0.1150**	**0.0525**
	Recall	0.1845	0.1320	**0.0610**
	F1	**0.1703**	**0.1229**	**0.0564**

5.4 Comparison with Long-Form Grounding Baselines

We compare MatchSense with TGB and Time-R1-style baselines. TGB represents context-extension methods using compact motion representations, while Time-R1 represents RL-based temporal grounding with chain-of-thought reasoning and verifiable temporal rewards.

Table 8: Efficiency of audio-guided pruning. MatchSense reduces the video duration processed by the video–language model while preserving most salient events. PM is "Processed Minutes".

Method	Video Min.	PM	Reduction	Speed
Dense Proc.	90.0	90.0	0.0%	1.0×
MatchSense	90.0	31.8	64.7%	2.8×

Table 9: Pruning quality and semantic density. MatchSense increases the density of useful content while retaining sufficient context for language reasoning.

Metric	Value
Timeline Retained	35.3%
FLOP Reduction	64.7%
Event Recall	83.0%
Global Precision	16.7%
Average Clip Precision	40.5%
Wall-clock Speedup	2.8×
Information Density Gain	2.35×

TGB achieves strong recall because it processes broader temporal context, but Precision@0.5 remains only 0.0450, showing that context extension alone does not solve the needle-in-a-haystack problem of ultra-long sports broadcasts. Time-R1 improves over TGB with 0.0725 mAP and 1.95s MBE, showing the value of RL and reasoning-guided localization. However, it lacks MatchSense's audio-guided structural prior and assignment-based multi-event reward construction. MatchSense improves mAP to 0.0942 and MBE to 1.54s. Implementation and adaptation details for TGB and Time-R1-style baselines are provided in Appendix A.10.

5.5 Reward Component Ablations

The reward ablations in Table 1 show that prosodic saliency and semantic alignment are both useful. Removing the temporally aligned prosody reward reduces mAP from 0.0815 to 0.0712 and increases MBE from 1.68s to 2.18s, showing that acoustic saliency anchors boundary refinement when coupled with temporal overlap. Removing the semantic reward reduces mAP to 0.0765 and lowers F1 at stricter tIoU thresholds, indicating weaker alignment between generated reasoning and event semantics.

5.6 Efficiency and Information Density

MatchSense retains only 35.3% of the original broadcast timeline, averaging 31.8 minutes per 90-minute match, while preserving 83% of salient events.

The retained timeline has 16.7% Global Precision, compared with approximately 7.1% for the original dense 90-minute broadcast. Thus, audio-guided pruning increases information density by 2.35× while reducing computation by 64.7%. Average Clip Precision of 40.5% shows that SOI clips contain substantial event-relevant content while preserving pre- and post-event context needed for coherent <think> reasoning.

5.7 Summary of Findings

The results support four conclusions. First, zero-shot video–language models are insufficient for full-length sports key moment identification. Second, structured SFT improves ontology alignment and output reliability but not precise temporal localization. Third, assignment-based GDPO improves mAP, Precision@0.5, and MBE by separately preserving temporal, semantic, acoustic, format, and penalty reward signals during policy optimization. Fourth, audio-guided pruning is essential for scalability, improving information density and reducing computation while preserving most salient events.

6 Conclusion

We presented **MatchSense**, an audio-guided language reasoning framework for key moment identification in long-form sports broadcasts. MatchSense uses prosodic saliency to extract Salient Object Intervals, applies structured SFT with <think> and <answer> outputs, and further optimizes variable-cardinality predictions through assignment-based *Group reward-Decoupled Normalization Policy Optimization* (GDPO). In this formulation, Hungarian matching produces event-level reward components, while GDPO separately normalizes temporal, semantic, acoustic, format, and penalty rewards before policy optimization.

Experiments on full-length soccer broadcasts show improvements in event localization, boundary precision, and efficiency over zero-shot, SFT-only, dense sliding-window, GRPO, and ablated variants. Overall, the results show that long-form sports understanding benefits from modality-specific pruning, structured language reasoning, and reward-decoupled multi-objective policy optimization rather than relying only on larger context windows or dense video processing.

7 Limitations

Several limitations remain. First, MatchSense relies on informative broadcast audio; prosodic saliency may be weaker with muted audio, delayed commentary, noisy mixing, or less expressive narration. Future work should combine audio with motion intensity, shot transitions, scoreboard changes, subtitles, or player-tracking cues. Second, false positives persist at strict tIoU thresholds because events such as fouls, shots, and cards have ambiguous buildup and delayed recognition, creating a mismatch between point-level annotations and extended semantic events.

Third, the BERTScore-based semantic reward may encourage commentary-consistent reasoning rather than purely visual understanding. Since the reference text is formed from the event label and ASR-aligned commentary, the model may receive semantic reward for producing language similar to the broadcast transcript even when visual grounding is incomplete. We reduce this risk by computing semantic reward only after Hungarian matching and combining it with temporal overlap and tIoU-scaled prosodic saliency, but disentangling visual reasoning from commentary mimicry remains open.

Fourth, GDPO-based policy optimization with large video–language models remains costly, requiring multiple generations, assignment-based multi-reward computation, reference-policy scoring, and Hungarian matching. We also do not claim statistical significance for the GDPO–GRPO gap. GDPO improves mAP from 8.15% to 9.42% and reduces MBE from 1.68s to 1.54s, but repeated runs with different seeds are needed to estimate variance.

Fifth, large-scale sports benchmarks for long-form temporal grounding and sparse event localization remain limited. Building such datasets requires temporally precise annotation of semantically meaningful events across extended broadcast videos, which is labor-intensive and difficult to automate reliably under ambiguous boundaries and multimodal context. Finally, experiments are limited to SoccerNet-v2 soccer broadcasts; other sports may require different event ontologies, saliency patterns, and annotation conventions. We discuss transferability beyond soccer in Appendix A.11.

8 Ethical Considerations

This work studies key moment identification in long-form sports broadcasts using SoccerNet-v2, accessed under the dataset's academic research agreement and usage terms. Our code is released via an anonymous repository for review and will be made publicly available under an open source license upon publication. SoccerNet-v2 is used strictly for academic research on sports video understanding, consistent with its intended use and access conditions. We do not redistribute the data, use it for commercial purposes, or apply it outside the research context outlined in the dataset license agreement. SoccerNet-v2 broadcast videos may contain identifiable players, referees, coaches, and commentators. We do not extract, store, or annotate any personal identity information. We use only timestamped event labels for research purposes. The framework is intended for research on sports video understanding, highlight generation, retrieval, summarization, and broadcast analysis. We do not use private user data, biometric identity labels, speaker identification, or personally sensitive annotations.

Dataset, privacy, and licensing. Our experiments rely on broadcast-style soccer videos and timestamped event annotations. Researchers using this framework should ensure that dataset access, storage, and redistribution comply with the relevant dataset licenses and copyright restrictions. Since sports broadcasts may contain identifiable players, referees, coaches, spectators, and commentators, deployment in non-public or amateur settings should obtain appropriate consent and avoid privacy-invasive uses.

Potential misuse and generated reasoning. Automated key moment identification can support beneficial applications such as highlight creation, video search, coaching support, and sports analytics. However, inaccurate or selectively edited outputs may misrepresent match narratives, player performance, or referee decisions. The generated `<think>` outputs are intermediate localization aids and should not be interpreted as complete causal explanations. We recommend using MatchSense as an assistive tool, with human verification for reporting, coaching, or decision-support applications.

Bias, coverage, and environmental impact. Our study focuses on SoccerNet-v2 soccer broadcasts, so model behavior may reflect dataset-

specific biases in camera style, commentary patterns, league coverage, annotation conventions, event frequency, and language. Performance may not transfer directly to other sports, lower-resource leagues, women's sports, amateur broadcasts, or videos with different commentary styles. Training and GDPO-based policy optimization with large video–language models also require substantial computation. MatchSense partially mitigates this by pruning the video timeline before expensive multimodal inference, but future work should further improve efficiency through smaller models, distillation, cached features, and more sample-efficient optimization.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. 2026. Temporal chain of thought: Long-video understanding by thinking in frames. *Advances in Neural Information Processing Systems*, 38:143018–143046.

Abdulrahman Baraka and Mohd Halim Mohd Noor. 2022. Weakly-supervised temporal action localization: a survey. *Neural Computing and Applications*, 34:1–21.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Alain De Cheveigné and Hideki Kawahara. 2002. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4508–4519.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721.

Silvio Giancola and Bernard Ghanem. 2021. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499.

Alan Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7:143 – 154.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. 2025. Revision-llm: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19012–19022.

Hadi Harb. 2009. Highlights detection in sports videos based on audio analysis.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and 1 others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681, New York, NY, USA. Association for Computing Machinery.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, pages 11846–11858. Curran Associates, Inc.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. 2021. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545.

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and 1 others. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy 2015*.

Michele Merler, Khoi-Nguyen C. Mac, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, Jinjun Xiong, Minh N. Do, John R. Smith, and Rogerio Schmidt Feris. 2019. Automatic curation of sports highlights using multimodal excitement features. *IEEE Transactions on Multimedia*, 21(5):1147–1160.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. pages 148–152.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, and 1 others. 2026. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364.

Ziyou Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang. 2003. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 3, pages III–401.

Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2678–2687.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Additional Implementation and Evaluation Details

This appendix provides additional implementation, dataset, prompting, optimization, and evaluation details for **MatchSense**. It complements the main

paper by expanding on audio-guided pruning, clip-level supervision, structured prompting, model selection, evaluation metrics, and GDPO-based multi-reward policy optimization.

A.1 Extended Motivation

A central bottleneck in long-form sports broadcast understanding is the mismatch between *sparse key moments* and *dense temporal input*. In a full soccer match, semantically consequential events such as goals, fouls, substitutions, cards, and shots occupy only a small fraction of the broadcast timeline. Sliding-window inference and dense proposal generation scale linearly with video duration and expose the model to large volumes of low-salience background play, worsening class imbalance and false positives (Lin et al., 2019; Zhang et al., 2022). As a result, many existing approaches restrict attention to short clips, rely on oracle proposals, or use heuristic pruning strategies that may not generalize to realistic broadcasts.

Policy optimization provides a mechanism for improving event-level localization beyond frame-level or token-level supervision (Yeung et al., 2016). However, applying it to long-form key moment identification is difficult because model outputs are set-valued: a clip may contain zero, one, or multiple events; event boundaries may overlap; and naive aggregated rewards can be unstable under false positives or missed detections (Baraka and Mohd Noor, 2022). MatchSense addresses this by combining audio-guided pruning, assignment-based reward construction, and reward-decoupled normalization across heterogeneous reward components.

Recent transformer-based video–language models and instruction-following systems improve multimodal temporal reasoning (Sun et al., 2019; Alayrac et al., 2022; Maaz et al., 2024), but most still treat time as uniformly searchable. In broadcast sports, commentator prosody, crowd reaction, acoustic onsets, and energy changes often indicate where important events occur. Prior multimedia analysis shows that audio energy, pitch variation, and perceived excitement correlate with salient sports moments (Hanjalic and Xu, 2005; Harb, 2009). MatchSense uses these cues as a structural prior that determines where language-guided video reasoning should be applied.

A.2 Extended Related Work

Temporal localization and sports action spotting. Temporal action localization detects and localizes events in untrimmed videos. Proposal-based methods such as Boundary-Matching Network explicitly model event boundaries (Lin et al., 2019), while anchor-free transformer models such as ActionFormer use multiscale temporal representations (Zhang et al., 2022). In sports, this is often framed as action spotting over full broadcasts. SoccerNet established this setting at scale, emphasizing sparse events, long duration, and sport-specific evaluation (Giancola et al., 2018; Deliege et al., 2021).

Language-grounded and multi-moment prediction. Video moment retrieval and temporal sentence grounding localize moments corresponding to natural-language queries. ActivityNet Captions, TVR, and MAD increased the linguistic and temporal complexity of grounding tasks (Krishna et al., 2017; Lei et al., 2020; Soldan et al., 2022). Realistic settings often require predicting multiple moments. QVHighlights models query relevance and saliency, while Moment-DETR casts moment retrieval as set prediction with bipartite matching (Lei et al., 2021). More broadly, DETR popularized Hungarian matching for end-to-end set prediction (Carion et al., 2020). MatchSense extends this set-based perspective to language-guided sports key moment identification and assignment-based multi-reward policy optimization.

Audio, prosody, and transcript signals. Audio has long been used for detecting highlights and affective moments in sports video. Acoustic energy, crowd noise, pitch variation, and commentator excitement provide strong saliency cues. Classical pitch estimators such as YIN remain widely used in audio-processing toolkits (De Cheveigné and Kawahara, 2002; McFee et al., 2015). ASR systems such as Whisper can provide timestamped transcripts for semantic alignment (Radford et al., 2023), although noisy or delayed broadcast speech motivates multimodal verification beyond transcript-only grounding (Koenecke et al., 2024). MatchSense uses prosodic saliency to structure the temporal search space before video–language reasoning.

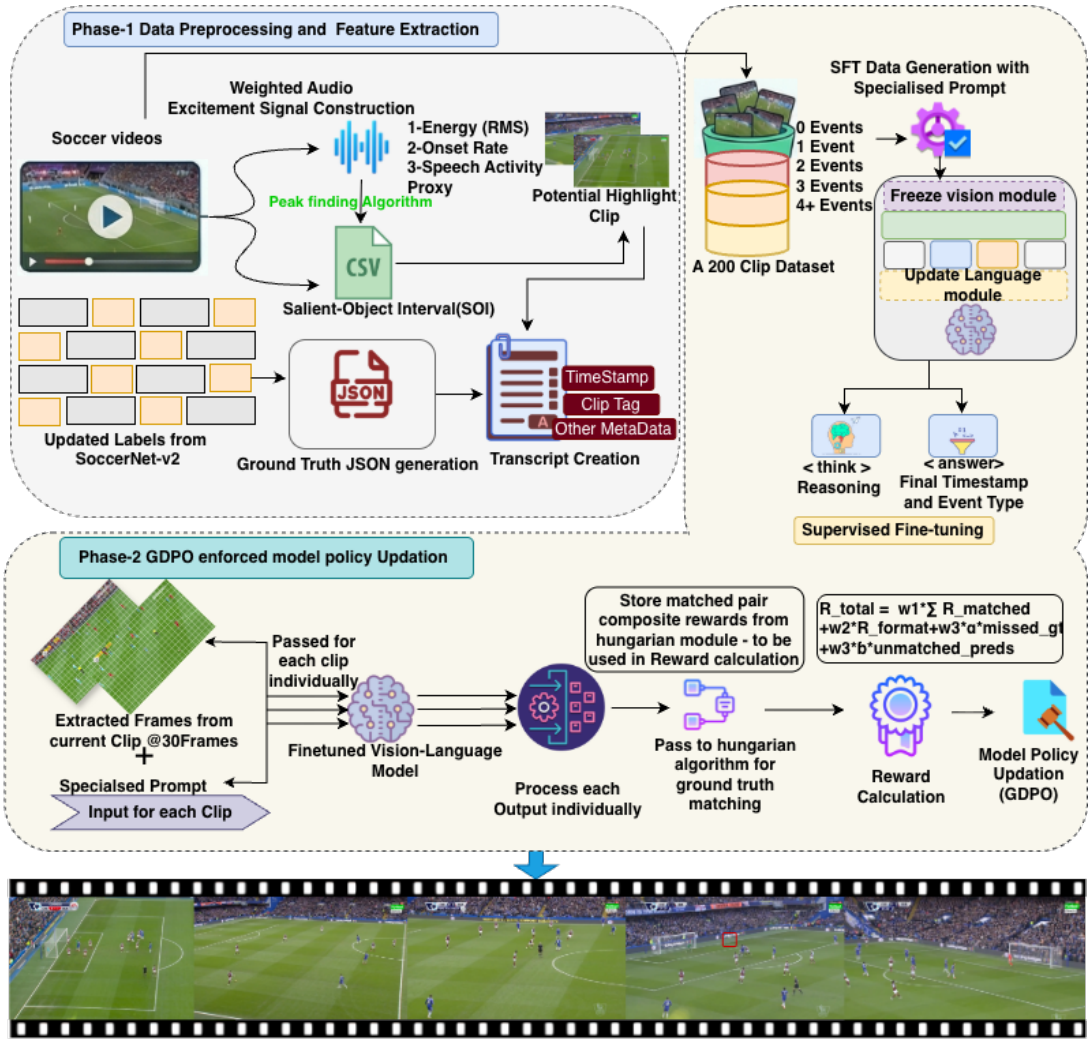

Figure 3: Detailed illustration of **MatchSense**. **Phase 1** performs audio-guided temporal pruning by constructing a weighted prosodic excitement signal from energy, onset rate, and speech activity, followed by peak detection to extract SOIs. **Phase 2** aligns SOIs with SoccerNet-v2 annotations to generate clip-level metadata, event labels, and transcript/context information. **Phase 3** uses SOI clips for structured SFT, where a video–language model with a frozen vision encoder generates `<think>` reasoning traces and `<answer>` key moment predictions. **Phase 4** applies assignment-based multi-reward policy optimization: candidate outputs are matched to ground-truth events through Hungarian assignment, converted into separate reward components, and optimized with GDPO through reward-decoupled normalization.

A.3 Detailed Pipeline Illustration

A.4 Adapting SoccerNet Annotations to Clip-Level Key Moment Identification

SoccerNet provides point-level annotations corresponding to nominal event timestamps. Following temporal localization practice, we expand each timestamp τ_i into a fixed interval:

$$[\tau_i - 7\,\mathrm{s},\ \tau_i + 7\,\mathrm{s}].$$

This captures annotation uncertainty, event buildup, and immediate aftermath, which are important in sports broadcasts where semantic evidence may precede or follow the nominal timestamp.

For each SOI clip spanning $[a_k, b_k]$, ground-truth intervals are transformed into clip-relative coordinates. If an expanded event interval is $[\tau_i^s, \tau_i^e]$, then its clip-relative interval is

$$[\tau_i^s - a_k,\ \tau_i^e - a_k].$$

During inference, predicted clip-relative intervals are mapped back to the global broadcast timeline by adding the SOI start time.

Annotation conversion and label filtering. During SoccerNet preprocessing, we exclude *Ball out of play* and *Throw-in*. These events occur frequently, dominate the temporal distribution,

and rarely correspond to semantically decisive or viewer-salient moments. Including them would intensify class imbalance and increase false positives without providing strong supervision for key moment identification.

Other contextual events, such as fouls, offsides, corners, substitutions, and clearances, are retained during training because they provide temporal structure and contextual supervision. Evaluation focuses on primary key moments that better reflect consequential sports events.

A.5 Training and Evaluation Prompt

The same instruction prompt is used across supervised fine-tuning, SFT dataset generation, and policy-optimization evaluation. This keeps training and evaluation distributions aligned and reduces prompt-induced artifacts.

The model outputs frame indices rather than absolute timestamps because it observes only the local clip. During post-processing, frame indices are mapped to seconds using the clip duration and frame sampling rate, and clip-relative timestamps are converted into global broadcast coordinates.

A.6 Fixed-Length Clip Sampling

Each clip is uniformly subsampled to a fixed budget of 30 frames to balance temporal coverage with context-length and memory constraints. All models and baselines use the same frame budget for fair comparison. For SFT, we additionally construct fixed-length 30-second clips and bucket them by event density, exposing the model to zero-event, single-event, and multi-event clips.

A.7 Efficiency of Audio-Guided Temporal Pruning

Audio-guided SOI extraction substantially reduces the search space. Using grid search over prosodic feature weights, the SOI stage achieves approximately **83% recall coverage** of ground-truth events. In practice, MatchSense retains 35.3% of the original timeline, corresponding to 31.8 minutes per 90-minute match. This gives a 64.7% FLOP reduction and a 2.8× wall-clock speedup compared with dense video processing. The retained timeline has 16.7% Global Precision and 40.5% Average Clip Precision. Compared with the original dense broadcast, where ground-truth event duration occupies approximately 7.1% of the timeline, SOIs provide a 2.35× increase in information density.

A.8 SOI Extraction Hyperparameters

We report the hyperparameters used for audio-guided Salient Object Interval (SOI) extraction to make the pruning stage reproducible. The prosodic excitement signal is constructed from RMS energy, localized onset-rate density, and an MFCC-derived speech-activity proxy:

$$E(t) = \alpha_1 E_{\text{energy}}(t) + \alpha_2 E_{\text{onset}}(t) + \alpha_3 E_{\text{speech}}(t),$$

where all features are min–max normalized to $[0, 1]$. We use

$$\alpha_1 = 0.50, \quad \alpha_2 = 0.30, \quad \alpha_3 = 0.20.$$

RMS energy is computed using a frame length of 2048 samples and hop length of 512 samples at a sampling rate of 16 kHz. Localized onset density is estimated from onset detections smoothed with a 1-second moving-average window. The speech-activity proxy is computed as the mean absolute MFCC magnitude across 13 coefficients.

After feature fusion, the excitement signal undergoes local baseline subtraction using a 2-second running mean, followed by re-normalization to $[0, 1]$. Temporal smoothing is then applied using a median filter with kernel size 9 (or 3 for very short sequences).

Candidate peaks are detected using `librosa` peak picking with

$$\texttt{delta_std_mult} = 3.2,$$

together with amplitude filtering at the 75th percentile and a minimum sustain duration of 1.5 seconds. Peaks are subsequently ranked by excitement magnitude, and non-maximum suppression is applied with a minimum temporal separation of 6 seconds. At most 40 SOIs are retained per video half.

Each selected peak t_p is expanded into a temporal interval

$$[t_p - \Delta_{\text{pre}},\ t_p + \Delta_{\text{post}}],$$

with

$$\Delta_{\text{pre}} = 20 \text{ seconds}, \quad \Delta_{\text{post}} = 5 \text{ seconds}.$$

Intervals whose boundaries fall within 12 seconds of each other are merged to form the final SOI set. These values were selected empirically to preserve high event recall while substantially reducing the amount of video processed by the video-language model.

Event Label Specification for Training and Evaluation

Evaluation Labels (Primary, Outcome-Altering Events):

- Penalty
- Indirect free-kick
- Direct free-kick
- Goal
- Shots on target
- Foul
- Yellow card
- Red card
- Yellow→Red card

Training Labels (Full Annotation Set):

- Penalty
- Kick-off
- Indirect free-kick
- Direct free-kick
- Corner
- Goal
- Substitution
- Offside
- Clearance
- Shots on target
- Shots off target
- Foul
- Yellow card
- Red card
- Yellow→Red card

Figure 4: Event label sets used for training and evaluation. Evaluation focuses on primary key moments, while training uses a broader annotation set to preserve contextual and causal supervision.

A.9 Choice of Base Vision–Language Model

We adopt Qwen2.5-VL-7B after empirical comparison with alternatives including Phi-3.5 and MiniCPM-V 2.6. In preliminary experiments, these alternatives often failed to follow the required structured output format, produced malformed responses, or hallucinated unsupported events. Some also could not process 30-second clips even with aggressive frame subsampling due to context limits. Qwen2.5-VL-7B offered the best balance of multimodal capacity, instruction-following reliability, context handling, and structured generation stability.

A.10 Long-Form Baseline Adaptation Details

We compare MatchSense with TGB- and Time-R1-style baselines as adapted long-form grounding baselines under the same SoccerNet-v2 key moment identification protocol. These baselines are not copied as reported leaderboard numbers from their original papers; instead, they are adapted to our clip construction, event ontology, and assignment-based evaluation.

The TGB-style baseline represents long-context temporal grounding with compact temporal representations. It processes longer temporal context than the base model and predicts event intervals

Prompt for Structured Football Key Moment Identification

Input: <video>

SYSTEM / ROLE: You are an expert Football Video Analyst and Sports Data Logger. Your task is to analyze the provided football match clip and output a structured log of key moments using a fixed reasoning-and-answer format.

TASK INSTRUCTIONS:

1. **Scan the Video:** Analyze the visual action at fine temporal resolution, approximately one second per frame.
2. **Generate <think>:** For each interval, write a concise description, fewer than six words, of what is visually happening, such as player running, passing, shooting, or referee signaling.
3. **Detect Key Moments:** Compare the observed action against the allowed key moment list.
4. **Generate <answer>:** If a key moment is detected, log it with its start and end frame indices. If multiple key moments occur, list all of them.
5. **No Events:** If no key moment from the list is detected, output exactly <answer></answer>.

ALLOWED KEY MOMENTS LIST:

- Penalty
- Kick-off
- Indirect free-kick
- Direct free-kick
- Corner
- Goal
- Substitution
- Offside
- Clearance
- Shots on target
- Shots off target
- Foul
- Yellow card
- Red card
- Yellow→Red card

OUTPUT FORMAT RULES:

- The output must contain exactly two tags: <think> and <answer>.
- The <think> tag must contain short frame intervals and concise descriptions.
- The <answer> tag must contain frame intervals and exact event labels from the allowed list.
- Frame intervals must use frame indices, e.g., 0-4 or 12-15.

Sample Output Format:
<think> 0-5: {desc_1}. 5-9: {desc_2}. 9-12: {desc_3}. 12-16: {desc_4}. 16-20: {desc_5}. </think>
<answer> 5-9: Event_1; 16-20: Event_2 </answer>

Figure 5: Structured prompt used for football key moment identification with intermediate language reasoning.

without audio-guided SOI pruning. The Time-R1-style baseline represents reinforcement-learning-based temporal grounding with reasoning traces and verifiable temporal rewards. For fair comparison, both baselines use the same underlying video–language backbone where applicable, the same 30-frame input budget, the same SoccerNet-v2 primary evaluation labels, and the same Hungarian matching evaluation protocol.

These adaptations are intended to test whether long-context grounding and RL-based temporal grounding are sufficient for sparse sports key moment identification. They should not be interpreted as exact reproductions of the original systems un-

der their native benchmarks or hyperparameter settings.

A.11 Transferability Beyond Soccer

Although all experiments are conducted on SoccerNet-v2 soccer broadcasts, the core components of MatchSense are not soccer-specific. The prosodic excitement signal is based on RMS energy, onset density, and speech-activity cues, which capture commentator arousal and crowd reaction patterns that occur across many televised sports, such as basketball dunks, cricket wickets, tennis break points, and American-football turnovers. These cues reflect broadcast production conventions rather than soccer-specific rules.

Similarly, the `<think>`–`<answer>` output format, Hungarian assignment-based reward construction, and GDPO optimization are defined over an abstract event ontology specified in natural language. Adapting MatchSense to another sport therefore mainly requires changing the allowed event label list and providing sport-specific SFT examples, rather than redesigning the architecture. The main sport-specific assumption is that broadcast audio correlates with visually salient events. We expect this assumption to hold broadly for professionally produced sports broadcasts, but the SOI thresholds and expansion windows may require recalibration for sports with different event densities, commentary styles, and crowd acoustics.

A.12 Mean Boundary Error at tIoU=0.5

Mean Boundary Error (MBE) is reported at tIoU 0.5 so that the metric reflects localization accuracy for successfully detected events, rather than being dominated by false positives at low overlap or near-perfect matches at high overlap. MBE complements precision, recall, F1, and mAP by measuring how tightly a detected key moment is localized around the ground-truth annotation.

A.13 Dataset Selection and Experimental Justification

We use SoccerNet-v2 because it directly matches sparse key moment identification in long-form sports broadcasts. Unlike Kinetics-400 or UCF101, which consist of short trimmed clips, SoccerNet provides full match broadcasts with sparse timestamped event annotations. Other datasets such as FineGym and MultiSports focus primarily on fine-grained action classification or spatiotemporal detection within shorter or pre-trimmed segments, making them less aligned with extracting discrete narrative events from continuous broadcast streams.

A.14 Extended Analysis of Hierarchical Temporal Pruning

Hierarchical pruning addresses the imbalance between informative and uninformative temporal regions in full-length sports broadcasts. Most of a soccer match consists of continuous background play, while key moments occupy only a small fraction of the timeline. Exhaustive processing is therefore computationally inefficient and statistically unstable.

Prosodic saliency provides an effective first-stage filter. By focusing video–language reasoning on acoustically salient intervals, MatchSense reduces irrelevant visual input and lowers the chance of hallucinated events during ordinary gameplay. Empirically, SOI extraction preserves high event recall while substantially reducing the processed timeline. Although audio prosody is especially effective for broadcast sports, the same framework can incorporate other lightweight cues, including motion intensity, subtitle timing, shot-change frequency, scoreboard changes, visual transitions, or player-tracking signals.

A.15 Extended Analysis of Assignment-Based Multi-Reward Optimization

Key moment identification naturally produces set-valued predictions with variable cardinality. A clip may contain multiple related events, such as a foul followed by a card or a shot followed by a corner. Greedy matching or fixed-order comparison can therefore produce unstable rewards.

MatchSense treats evaluation and reward construction as an optimal bipartite matching problem. Hungarian matching produces a one-to-one correspondence between predicted and ground-truth key moments, making temporal and semantic rewards permutation-invariant and robust to prediction order. The matched outputs are then decomposed into separate reward components, including temporal overlap, semantic alignment, temporally aligned acoustic saliency, output-format validity, and penalties for unmatched or missed events.

This decomposition is important because the reward components are heterogeneous: format validity, tIoU, semantic alignment, acoustic saliency, and missed-event penalties have different scales and optimization dynamics. Instead of collaps-

Table 10: Comparison of sports and action recognition datasets. SoccerNet is uniquely aligned with long-form key moment identification because it provides full broadcasts with sparse timestamped events.

Dataset	Video Format	Avg. Duration	Annotation Type	Event Density	Primary Task
SoccerNet (Ours)	**Untrimmed Broadcast**	**∼90 min**	**Timestamp**	**Sparse**	**Key Moment Extraction**
Kinetics-400 (Kay et al., 2017)	Trimmed Clips	∼10 sec	Action Class	Dense	Action Classification
FineGym (Shao et al., 2020)	Trimmed Routines	∼10 sec	Hierarchical Segments	Dense	Fine-grained Analysis
MultiSports (Li et al., 2021)	Trimmed Clips	∼20 sec	Spatiotemporal BBox	Dense	Spatiotemporal Detection
UCF101 (Soomro et al., 2012)	Trimmed Clips	∼7 sec	Action Class	Dense	Action Classification

ing them into a single scalar before normalization, MatchSense uses GDPO to normalize each reward component separately within a rollout group before aggregation. This preserves distinctions among reward dimensions and provides a more informative policy-update signal. Empirically, assignment-based GDPO improves both event-level accuracy and boundary precision, indicating that gains are not caused merely by generating more predictions.

A.16 Annotation Assumptions

MatchSense assumes access to event annotations that can be expanded into short temporal intervals. This is standard in sports analytics and temporal localization benchmarks, but may not hold in domains with noisy, weakly aligned, or inconsistently defined annotations. Point-level labels are also imperfect for temporally extended events. For example, a foul may include buildup, contact, referee response, and player reactions. Fixed-window expansion partially addresses this issue, but future work should explore adaptive event windows, uncertainty-aware labels, and learning directly from point-level supervision.

A.17 Computational Cost of Policy Optimization

Although audio-guided pruning reduces the number of clips processed by the video–language model, policy optimization remains computationally expensive. Training requires repeated candidate generation, assignment-based reward computation, reference-policy evaluation, and rollout-group normalization. Future work could explore more sample-efficient policy objectives, offline construction of rollout groups, batched reward computation, reward-model distillation, or lightweight student models trained from MatchSense outputs.

A.18 Structured Output Constraints

The fixed `<think>` and `<answer>` format simplifies parsing, evaluation, and reward computation. It also provides an interpretable reasoning trace that can be aligned with commentary-derived semantic signals. However, this structure may limit expressiveness in tasks requiring richer narrative explanation, confidence estimation, or open-ended summarization. Future work could extend the output representation to include confidence scores, hierarchical event descriptions, or natural-language explanations while preserving reliable parsing.

A.19 Qualitative Analysis of Language-Guided Grounding

We analyze representative `<think>`–`<answer>` outputs to understand how structured language reasoning affects key moment identification. All examples follow the same 30-frame budget used in our pipeline, corresponding to uniform 1 fps sampling over a 30-second clip. The generated descriptions also follow the prompt constraint that each local description should contain fewer than six words.

Correct multi-event reasoning. Figure 6 shows a successful case where the model converts continuous play into a causal sequence before predicting event intervals. The reasoning trace describes the progression from a tackle to referee intervention and finally to a yellow card. This supports the role of intermediate language as a semantic bridge between visual evidence and event-level prediction.

Modality-induced hallucination. Figure 7 shows a false-positive case caused by strong acoustic saliency. The SOI is useful because the crowd reaction indicates a salient moment, but the model over-interprets the audio as visual confirmation of a goal. It therefore produces a plausible but unsupported reasoning trace and predicts `Goal`, although the ground truth is `Shots off target`. This illustrates a key limitation of audio-guided pruning: paralinguistic cues help identify candidate regions, but they can bias the video–language model when audio excitement conflicts with fine-grained visual evidence.

Case A: Correct Multi-Event Reasoning

Ground Truth:
8–12: Foul; 16–20: Yellow card

Model Output:
<think>
0–8: players running with ball.
8–12: defender tackles attacker aggressively.
12–16: referee blows whistle running.
16–20: referee shows yellow card.
20–29: players setup free-kick.
</think>
<answer>
8–12: Foul; 16–20: Yellow card
</answer>

Figure 6: Successful structured reasoning example. The model identifies a causal chain from tackle to whistle to yellow card, and predicts both key moments with correct intervals.

Case B: Modality-Induced Hallucination

Ground Truth:
10–15: Shots off target

Model Output:
<think>
0–10: attacking team passing outside box.
10–15: player shoots ball towards net.
15–22: ball hits net crowd cheering.
22–29: players celebrating near corner.
</think>
<answer>
10–15: Goal
</answer>

Figure 7: False-positive example caused by modality-induced hallucination. Strong crowd reaction biases the generated reasoning toward a goal-like interpretation despite the ground truth being a shot off target.

Temporal quantization bottleneck. Figure 8 shows a missed-event case. The substitution occurs in the background while play is stopped, but uniform 30-frame sampling does not provide a distinct visual or linguistic anchor for the event. As a result, the model summarizes the interval as a generic camera movement and emits an empty <answer> block. This suggests that structured reasoning improves localization for visually salient macro-events, but remains limited for brief or peripheral events that lack sustained camera focus.

Analysis. These examples reveal three behaviors. First, successful predictions often exhibit *causal semantic anchoring*: the <think> trace decomposes a sports event into buildup, trigger, and consequence before the <answer> prediction. In our GDPO formulation, this behavior is encouraged through the semantic reward R_{sem}, which uses BERTScore alignment between generated reasoning and reference event/commentary text. Second, false positives often arise from *modality-induced hallucination*, where strong acoustic excitement causes the model to justify an incorrect event label through unsupported language. Third, false negatives often reflect a *temporal quantization bottleneck*: brief or off-ball events may be missed when uniform frame sampling fails to allocate a clear visual frame to the event.

Case C: Missed Event from Temporal Quantization

Ground Truth:
12–18: Substitution

Model Output:
<think>
0–10: play stopped for throw-in.
10–20: camera pans across field.
20–29: player throws ball into play.
</think>
<answer></answer>

Figure 8: False-negative example caused by temporal quantization. A brief off-ball substitution is compressed into a generic camera-pan description, leading to an empty prediction.

Implications. The qualitative analysis supports the main quantitative findings: structured reasoning improves interpretability and event-level localization, but audio-guided language reasoning must be paired with stronger visual verification and more adaptive temporal sampling. Future work should explore cross-modal consistency checks and dynamic frame allocation to reduce hallucinations and missed micro-events.

B Discussion

Our results show that key moment identification in long-form sports broadcasts benefits from combining audio-guided temporal pruning, structured language reasoning, and assignment-based GDPO. This combination addresses three central difficulties: sparse consequential events, the need to translate continuous multimodal evidence into discrete event labels and boundaries, and variable-cardinality key moment prediction under heterogeneous reward signals.

B.1 Audio-Guided Pruning and Language Reasoning

The audio-guided SOI stage provides a structural prior for long-form broadcast understanding. Dense-window evaluation exposes the model to large amounts of low-value background play, increasing false positives and weakening boundary precision. In contrast, audio-guided pruning concentrates the model's reasoning budget on regions where commentator excitement, crowd reaction, and acoustic transitions indicate likely semantic events.

The structured `<think>`–`<answer>` output format further supports localization by turning event detection into a language-guided reasoning problem. The model first describes local play dynamics and then predicts event labels and intervals. The no-CoT ablations show that this reasoning trace improves localization quality and provides the textual surface needed for semantic reward computation.

B.2 Assignment-Based GDPO

The assignment-based reward addresses the set-valued structure of sports key moment identification. Hungarian matching prevents duplicate credit, supports variable prediction cardinality, and stabilizes reward computation. SFT aligns the model with the event ontology and output format, but it does not explicitly optimize temporal overlap.

By contrast, assignment-based GDPO first derives separate reward components from matched predictions and then applies group reward-decoupled normalization. Temporal overlap, semantic alignment, temporally aligned acoustic saliency, format validity, and missed-event penalties are normalized separately within each rollout group before being aggregated into the final advantage. This avoids over-compressing heterogeneous rewards into a single normalized scalar and leads to higher mAP, higher Precision@0.5, and lower MBE than GRPO.

B.3 Future Directions

Several extensions are promising. First, additional lightweight saliency signals could improve robustness when audio is weak, delayed, or noisy. Second, theoretical analysis of assignment-based multi-reward normalization could strengthen the learning formulation. Third, applying MatchSense to sports such as cricket, basketball, or tennis would test whether audio-guided language reasoning generalizes across event structures and broadcast styles. Finally, extending the framework to non-sports long-form videos would assess its broader applicability for multimodal key event identification.

Summary. Overall, audio-guided pruning, structured language reasoning, and assignment-based GDPO form a coherent approach to key moment identification under extreme temporal sparsity. While this work focuses on soccer broadcasts, the framework is modular and can be adapted to other long-form multimodal domains.

B.4 Group Relative Policy Optimization

We evaluate *Group Relative Policy Optimization* (GRPO) as a comparison variant for assignment-based multi-reward policy optimization. GRPO aggregates the reward components into a single scalar and then normalizes the summed reward within a group of sampled outputs.

Problem setup. Let $\pi_\theta(y \mid x)$ denote the current policy, where x is a video-conditioned prompt and y is a structured output containing a reasoning trace and one or more predicted key moment intervals. For each input x, we sample a group of K candidate outputs:

$$\{y_1, \ldots, y_K\} \sim \pi_\theta(\cdot \mid x).$$

For each candidate y_j, Hungarian matching produces reward components:

$$\mathbf{R}^{(j)} = \left[R^{(j)}_{\text{fmt}}, R^{(j)}_{\text{tIoU}}, R^{(j)}_{\text{sem}}, R^{(j)}_{\text{aud}}, R^{(j)}_{\text{pen}}\right].$$

GRPO first aggregates them into a scalar reward:

$$R^{(j)}_{\text{sum}} = \lambda_f R^{(j)}_{\text{fmt}} + \lambda_t R^{(j)}_{\text{tIoU}} + \lambda_s R^{(j)}_{\text{sem}} + \lambda_a R^{(j)}_{\text{aud}} + \lambda_p R^{(j)}_{\text{pen}}.$$

Relative advantage estimation. GRPO computes a group-relative advantage from the summed reward:

$$A^{(j)}_{\text{GRPO}} = \frac{R^{(j)}_{\text{sum}} - \mu_{\text{sum}}}{\sigma_{\text{sum}} + \epsilon},$$

where μ_{sum} and σ_{sum} are the mean and standard deviation of $\{R^{(1)}_{\text{sum}}, \ldots, R^{(K)}_{\text{sum}}\}$. This is simple and effective for single-reward settings, but in multi-reward optimization it can collapse distinct reward combinations into identical advantage values.

GRPO objective. The policy is optimized with the standard clipped policy objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{1}{|y_j|}\sum_{t=1}^{|y_j|}\min\Big(\rho_{j,t}(\theta)A^{(j)}_{\text{GRPO}}, \text{clip}(\rho_{j,t}(\theta), 1-\epsilon_c, 1+\epsilon_c)A^{(j)}_{\text{GRPO}}\Big)\right]$$

where

$$\rho_{j,t}(\theta) = \frac{\pi_\theta(y_{j,t} \mid x, y_{j,<t})}{\pi_{\theta_{\text{old}}}(y_{j,t} \mid x, y_{j,<t})}.$$

GRPO is reported as a comparison variant, while GDPO is the main MatchSense optimization method.

B.5 Group reward-Decoupled Normalization Policy Optimization

We use *Group reward-Decoupled Normalization Policy Optimization* (GDPO) as the main policy-optimization objective for MatchSense. GDPO is designed for multi-reward RL settings where different reward components capture distinct objectives. Instead of summing all rewards first and then applying group normalization, GDPO normalizes each reward component separately within the rollout group.

Scope of GDPO adaptation. GDPO is not introduced as a new optimization algorithm in this work. Our contribution is to adapt it to long-form sports key moment identification by coupling reward-decoupled normalization with Hungarian event matching and multimodal reward components derived from temporal overlap, BERTScore-based semantic alignment, temporally coupled acoustic saliency, output-format validity, and cardinality penalties.

Reward-decoupled normalization. Given the reward vector

$$\mathbf{R}^{(j)} = \left[R^{(j)}_{\text{fmt}}, R^{(j)}_{\text{tIoU}}, R^{(j)}_{\text{sem}}, R^{(j)}_{\text{aud}}, R^{(j)}_{\text{pen}}\right],$$

GDPO computes a normalized advantage for each reward component q:

$$A^{(j)}_q = \frac{R^{(j)}_q - \mu_q}{\sigma_q + \epsilon},$$

where μ_q and σ_q are the mean and standard deviation of reward component q across the sampled group.

Multi-reward advantage aggregation. The component-wise advantages are then combined using reward weights: $A^{(j)}_{\text{sum}} = \lambda_f A^{(j)}_{\text{fmt}} + \lambda_t A^{(j)}_{\text{tIoU}} + \lambda_s A^{(j)}_{\text{sem}} + \lambda_a A^{(j)}_{\text{aud}} + \lambda_p A^{(j)}_{\text{pen}}$. To stabilize numerical scale across mini-batches, we apply batch-wise advantage normalization:

$$\widehat{A}^{(j)}_{\text{sum}} = \frac{A^{(j)}_{\text{sum}} - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}} + \epsilon}.$$

GDPO objective. The policy is optimized using the clipped objective:

$$\mathcal{J}_{\text{GDPO}}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{1}{|y_j|}\sum_{t=1}^{|y_j|}\min\Big(\rho_{j,t}(\theta)\widehat{A}^{(j)}_{\text{sum}}, \text{clip}(\rho_{j,t}(\theta), 1-\epsilon_c, 1+\epsilon_c)\widehat{A}^{(j)}_{\text{sum}}\Big)\right].$$

This objective preserves the contribution of each reward dimension while retaining the stability of group-based policy optimization.

Relation to assignment-based rewards. The reward components used by GDPO are computed after Hungarian matching between predicted and ground-truth key moments. Thus, temporal overlap, semantic alignment, acoustic saliency, format validity, and penalty signals are all assignment-consistent. GDPO then prevents these heterogeneous signals from being collapsed into a single normalized reward before policy optimization.

Comparison with GRPO. GRPO and GDPO differ in how they normalize multi-reward signals. GRPO normalizes the summed reward, which can obscure distinctions among reward combinations. GDPO normalizes each reward component separately before aggregation, preserving more fine-grained reward information. We evaluate both, but GDPO is the main MatchSense method because it obtains the best empirical performance.

B.6 Dual-Group Reward Architecture and Hyperparameter Values

MatchSense uses a dual-group reward architecture to separate event-level assignment from global sequence control. This separation is important because key moment identification is a variable-cardinality set prediction problem: the model must both localize matched events accurately and decide how many events to emit.

Figure 9 illustrates how the assignment-based reward is computed before GDPO normalization.

Table 11: Comparison of policy optimization methods used for reward-based learning. The table highlights differences in value-function use, reward normalization, group-wise learning, and suitability for structured key moment prediction.

Method	Value Function	Reward Normalization	Group-wise	Key Characteristic
PPO	✓	Scalar advantage with learned critic	✗	Clipped policy gradients with a learned value model for stable reinforcement learning updates.
GRPO	✗	Group-normalized summed reward	✓	No critic model; optimization may collapse heterogeneous reward signals into a single normalized objective.
GDPO	✗	Group-normalized per-reward components	✓	Reward-decoupled normalization designed for heterogeneous multi-reward reinforcement learning settings.

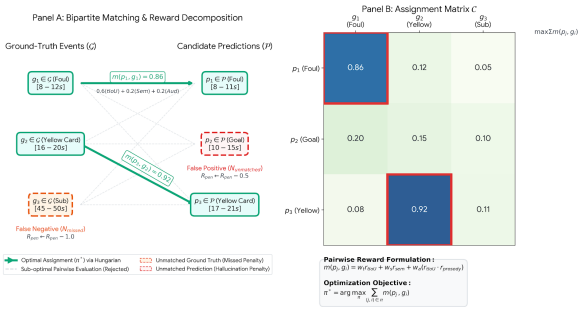

Figure 9: Assignment-based multi-reward matching for GDPO. **Panel A** illustrates bipartite matching between ground-truth events $\mathcal{G}$ and candidate predictions $\mathcal{P}$. Matched pairs receive a positive composite matching score based on temporal overlap, semantic alignment, and temporally coupled acoustic saliency. Unmatched ground-truth events incur a missed-event penalty of -1.0, while unmatched predictions incur a hallucination penalty of -0.5. **Panel B** shows the corresponding positive matching-score matrix, where each displayed entry is $m(p_j, g_i)$. For implementation with a minimization-based Hungarian solver, we pass the negated cost matrix $C_{j,i} = -m(p_j, g_i)$, which is equivalent to maximizing the global matching score before reward-decoupled GDPO normalization.

Because key moment identification produces set-valued outputs with variable cardinality, the number of predicted events need not match the number of ground-truth events. Directly comparing predictions in generation order would therefore make the reward sensitive to arbitrary ordering and duplicate predictions.

We instead construct a bipartite matching problem between the ground-truth set $\mathcal{G}$ and the prediction set $\mathcal{P}$. Each candidate pair (p_j, g_i) receives the positive composite matching score $m(p_j, g_i) = w_t r_{\text{tIoU}}(p_j, g_i) + w_s r_{\text{sem}}(p_j, g_i) + w_a\big(r_{\text{tIoU}}(p_j, g_i) \cdot r_{\text{prosody}}(g_i)\big)$, with $(w_t, w_s, w_a) = (0.6, 0.2, 0.2)$. The optimal assignment is therefore

$$\pi^* = \arg\max_{\pi} \sum_{(j,i)\in\pi} m(p_j, g_i).$$

In implementation, because the Hungarian solver expects a minimization objective, we pass the equivalent negated cost matrix

$$C_{j,i} = -m(p_j, g_i).$$

Thus, the positive values shown in Figure 9 are matching scores for interpretability, while the internal solver cost uses the negative sign.

Unmatched elements are handled after assignment. A missed ground-truth event receives a stronger penalty because sparse sports key moments are consequential and should not be omitted. An unmatched prediction receives a milder hallucination penalty to discourage false positives without forcing the model into overly conservative empty outputs:

$$R_{\text{pen}} = -0.5N_{\text{unmatched}} - 1.0N_{\text{missed}}.$$

After matching, the temporal, semantic, acoustic, format, and penalty components are passed separately to GDPO. This is important because GDPO normalizes each reward component independently before aggregation, preserving fine-grained boundary, semantic, and acoustic learning signals that may otherwise collapse under a single summed reward.

Group 1: assignment-based matching rewards. The first group is evaluated inside the Hungarian matching step and determines the one-to-one assignment between predicted key moments and ground-truth events. For a predicted event p_j

and ground-truth event g_i, we define the pairwise matching score as

$$m(p_j, g_i) = w_t r_{\text{tIoU}}(p_j, g_i) + w_s r_{\text{sem}}(p_j, g_i) + w_a\big(r_{\text{tIoU}}(p_j, g_i) \cdot r_{\text{prosody}}(g_i)\big),$$

with

$$w_t = 0.6, w_s = 0.2, w_a = 0.2, \qquad w_t + w_s + w_a = 1.$$

The temporal term receives the largest weight because fine-grained coordinate regression is difficult for autoregressive video–language models and should dominate event matching. The semantic term measures BERTScore alignment between the generated <think> trace and the reference event/commentary text. The acoustic term rewards alignment with broadcast excitement, but only when the prediction temporally overlaps the event.

Temporally coupled prosody reward. We use the product

$$r_{\text{tIoU}}(p_j, g_i) \cdot r_{\text{prosody}}(g_i)$$

rather than adding $r_{\text{prosody}}(g_i)$ as an independent term. A standalone prosody score depends only on the ground-truth event and is therefore constant across a matching-matrix column. Such a constant does not help distinguish competing predictions for the same event during Hungarian assignment. Multiplying prosody by tIoU couples acoustic saliency with temporal precision: a prediction receives a stronger acoustic reward only if it overlaps the prosodically salient event. Thus, the acoustic prior supports localization instead of acting as a flat event-level bias.

Group 2: global and syntactic sequence rewards. The second group is evaluated after Hungarian matching and controls output validity and prediction cardinality: $R_{\text{fmt}} = r_{\text{format}}(y)$, $R_{\text{pen}} = -\lambda_{\text{FP}} N_{\text{unmatched}} - \lambda_{\text{FN}} N_{\text{missed}}$. We use a strict binary format reward $r_{\text{format}}(y) \in \{0, 1\}$ that checks whether the model follows the required <think> and <answer> syntax. For cardinality control, we use asymmetric penalties:

$$\lambda_{\text{FP}} = 0.5, \qquad \lambda_{\text{FN}} = 1.0.$$

The missed-event penalty is harsher because sparse sports key moments are rare and recall is important. The false-positive penalty is still milder than the missed-event penalty, avoiding collapse into overly conservative empty predictions while discouraging hallucinated extra events.

Preventing multimodal reward hacking. Evaluating temporal, semantic, and acoustic signals jointly inside the Hungarian matching step reduces reward hacking. A prediction cannot receive a high semantic score alone unless it is also selected through assignment against a temporally plausible ground-truth event. Likewise, acoustic saliency contributes only through the tIoU-scaled prosody product. This design encourages the model to produce predictions that are simultaneously localized, semantically consistent, and acoustically grounded.

Why GDPO is needed. The resulting reward vector contains heterogeneous components:

$$\mathbf{R}^{(j)} = \left[R^{(j)}_{\text{fmt}}, R^{(j)}_{\text{tIoU}}, R^{(j)}_{\text{sem}}, R^{(j)}_{\text{aud}}, R^{(j)}_{\text{pen}}\right].$$

These signals have different scales and learning dynamics. Format validity is often solved early and may lose variance, while temporal overlap and semantic alignment remain continuous and difficult. Standard GRPO first sums the reward components and then normalizes the aggregate, which can collapse distinct reward profiles into similar advantages. GDPO instead normalizes each reward component separately within the rollout group before summation. This preserves fine-grained temporal, semantic, acoustic, format, and penalty signals during policy optimization.

B.7 Training Reward Dynamics of GDPO and GRPO

Figure 10 compares the overall aggregated reward per SOI clip during one epoch of policy optimization. Both GRPO and GDPO begin with strongly negative rewards, around -2.6, and briefly dip to approximately -2.76. This early penalty-dominated regime occurs because the model frequently misses ground-truth events, produces unmatched predictions, and has low temporal-overlap and semantic-alignment scores.

Preventing reward-signal collapse. The reward contains heterogeneous signals: temporal overlap with weight 0.6, semantic alignment with weight 0.2, temporally coupled acoustic saliency with weight 0.2, format validity, and asymmetric cardinality penalties for missed and hallucinated events. GRPO aggregates these components before group normalization, so high-variance temporal rewards, low-variance format rewards, and discrete penalties can obscure one another. This causes the smoothed GRPO reward to improve more slowly

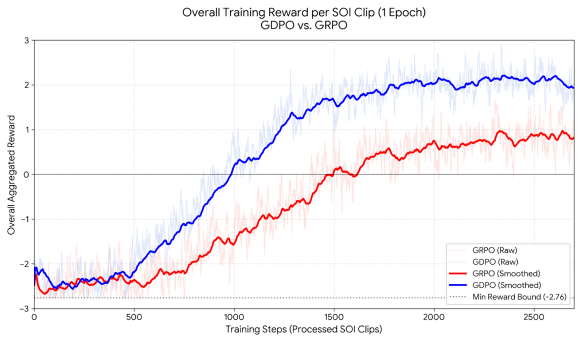

Figure 10: Overall aggregated training reward per SOI clip for GDPO and GRPO over one epoch. Both methods start in a penalty-dominated regime, but GDPO rises faster and reaches a higher plateau because each reward component is normalized separately before aggregation.

and plateau near 0.85. GDPO instead normalizes each reward component separately within the rollout group before aggregation, allowing temporal, semantic, acoustic, format, and penalty signals to remain distinguishable. This leads to a substantially higher reward trajectory, plateauing above 2.1.

Continuous reward versus discrete evaluation. The large training-reward gap between GDPO and GRPO does not translate linearly into mAP because policy optimization uses a continuous proxy reward, whereas evaluation uses hard tIoU thresholds. Small boundary refinements can increase the training reward continuously, but a prediction receives evaluation credit only after crossing the required tIoU threshold. Thus, large gains in the continuous reward space can correspond to more modest improvements in thresholded metrics such as mAP and Precision@0.5.

Boundary refinement. The practical benefit of GDPO is most visible in boundary quality. By preserving the temporal reward signal during optimization, GDPO reduces Mean Boundary Error from 1.68s under GRPO to 1.54s. This helps borderline predictions cross the tIoU 0.5 threshold, contributing to the final gains in mAP (9.42%) and Precision@0.5 (11.50%).

B.8 Use of AI Assitants

We used AI coding assistants, including GitHub Copilot and/or OpenAI ChatGPT, during the implementation of MatchSense. Specifically, AI assistance was used for debugging and the implementation of standard utilities such as data loading, preprocessing pipelines, evaluation scripts, English editing, and grammar refinement. All research contributions, experimental design, model architecture decisions, reward formulations, and conclusions are solely the authors' own work. All AI-generated code was manually reviewed and verified by the authors before use.